维语网页中 *n*-gram 模型结合类不平衡 SVM 的不良文本过滤方法 *

如先姑力·阿布都热西提 [1]，亚森·艾则孜 [1†]，郭文强 [2]

(1. 新疆警察学院 信息安全工程系，乌鲁木齐 830013; 2. 新疆财经大学 计算机科学与工程学院，乌鲁木齐 830013)

摘　要：随着新疆地区网络的建设发展，产生了大量维吾尔语网页。为了构建健康网络环境，提出了一种结合 n-gram 统计模型和类不平衡支持向量机（SVM）分类器的维语文本过滤方法。首先，将网页文本进行预处理操作，通过 n-gram 统计模型来初步提取词干；然后，对词干进行语义分析，将具有相似含义的词干聚合为一类，以此降低词干维度；最后，在传统 SVM 中引入一个控制超平面之间距离的参数，构建一种类不平衡 SVM，使其能够很好地分类具有非线性不可分和不平衡性的维吾尔语文本。实验结果表明，该方法能够准确分类出不良文本，且具有较短的分类时间。

关键词：维吾尔语网页；不良文本过滤；n-gram 词干提取；类不平衡 SVM

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.07.0410

Reactionary text filtering method based on *n*-gram and class-unbalanced SVM for uyghur webpages

Ruxianguli·Abudurexiti[1], Yasen Aizezi[1†], Guo Wenqiang[2]

(1. *Dept. of Information Security Engineering Xinjiang Police College*, *Urumqi* 830013, *China*; 2. *School of Computer Science & Engineering Xinjiang University of Finance & Economic*, *Urumqi* 830013, *China*)

Abstract: Along with the construction and development of the network in Xinjiang, a large number of Uyghur webpages have been produced. In order to construct a healthy network environment, this paper proposed a Uyghur text filtering method combining n-gram statistical model and class-unbalanced support vector machine (SVM) classifier. Firstly, it preprocessed the webpage text, and extracted the stem initially by the N-gram statistical model. Then, it carried out the semantic analysis of the stems, and aggregated the stems with similar meanings into one class, thereby reducing the stem dimension. Finally, it introduced a parameter that controls the distance between hyperplanes in the traditional SVM, and constructed a class-unbalanced SVM to classify Uyghur texts with nonlinear indivisibility and imbalance. The experimental results show that the method can accurately classify bad texts and has a shorter classification time.

Key words: Uyghur webpage; reactionary text filtering; n-gram stem extraction; class-unbalanced SVM

0　引言

由于互联网的迅速发展和普及，产生了大量不同类型的短文本，例如网页论坛、推文、新闻提要、书籍、电影概要等。对短文本进行分类对于网络信息过滤非常重要[1]，将分类为存在毒品、色情等不健康文本进行过滤能够净化网络环境，有助于维护社会稳定。短文本通常是非结构化的，并采用简短对话的形式，由多个短句组成。由于短文本具有稀疏特征向量和类别不平衡性，因此不能使用传统分类技术来对其进行高准确性分类[2,3]。不平衡数据集是指其中不同类别的样本数量不均。一个包含很多样本的类被称为“多数类”，相反，包含很少样本的类被称为“少数类”。在对不平衡数据集进行分类时，分类器往往对多数类达到较高准确性，但对少数类的准确率较低[4]。

近些年，随着新疆经济和教育的发展，产生了很多维吾尔语网站。对维语网站中的不良文本信息进行过滤对新疆的稳定和健康发展具有重要意义。目前，对于维吾尔语文本的分类和过滤方法研究较少，主要为新疆大学。例如，文献[5]通过使用纯粹的统计学方法，仅依赖于单词的 N-gram 来进行分类。文献[6]中应用了一种监督方法，使用最大熵分类器将文档分类为已知类别，以及使用一种无监督学习方法，将未标记文档进行分组，其特征向量由原始单词和其 N-gram 组成。文献[7]中使用了一种 K 最近邻（KNN）分类器，并结合了三种不同的距离

收稿日期：2018-07-04；**修回日期：**2018-08-28　　**基金项目：**国家自然科学基金资助项目（61762086）；新疆维吾尔自治区高校科研计划项目（XJEDU2017M046）；国家社会科学基金资助项目（13CFX055）

作者简介：如先姑力·阿布都热西提（1976-），女（维吾尔族），新疆喀什人，副教授，硕士，主要研究方向为文本分类、信息安全等；亚森·艾则孜（1975-），男（维吾尔族）（通信作者），新疆库车人，国家电子数据司法鉴定员，教授，硕士，主要研究方向为数字取证、自然语言处理等；郭文强（1975-），男（锡伯族），新疆伊宁人，教授，博士，主要研究方向为信息安全.

度量（余弦，欧氏和 Jaccard）。文献[8]提出使用 χ^2 方法进行特征提取，并采用了支持向量机（support vector machine，SVM）作为分类器。使用 χ^2 统计来选择特征，如果特征和类是独立的，则 χ^2 的值为零。这种方法具有较高的特征维数空间，因为文档空间矢量是稀疏的，很少有特征是不相关的。

另外，对于不平衡数据集的分类，传统 SVM 并不能得到很好的结果。为此，一些学者对其进行了改进，例如，文献[9]对 SVM 中多数和少数类使用不同的损失函数（即使用 L_2 范数的平方而不是 L_1 范数），来惩罚少数数据样本的错误分类。文献[10]试图通过向少数类样本的支持向量中引入校正因子来校正分类器学习的偏移，以减少 SVM 模型的偏差。文献[11]中提出模糊支持向量机（FSVM）作为学习工具。这些技术需要微调用户定义的参数，具有高度复杂性。文献[12]中采用了少类样本合成过采样技术（SMOTE），是一种较新的非均衡数据集学习办法，通过对少数类样本的人工合成来提高其比例，以平衡样本之间的差异。但是，SMOTE 需要对许多用户定义的参数进行微调，获得一组合适参数较为困难。文献[13]中提出了一种称为 MINSVM 的改进型 SVM 方法，其通过删减一部分多数类样本，并为少数类样本提供更大的权重，使它们比多数类更受关注，致使产生的超平面应尽可能地接近多数类别。然而，其删减样本的操作一定程度上必然会影响学习效果。

为此，本文结合了 n-gram 统计模型和类不平衡 SVM 分类器，提出了一个新的框架来分类网页中的短维吾尔语文本。实验结果表明了提出方法能够有效的分类不良文本，为网页不良信息过滤提供了良好基础。

1　提出的文本过滤框架

维吾尔语是以阿拉伯字母为基础的文字，具有高度的黏着性。维吾尔字母共有 32 个，字母的形式具有多样性，通常包含 4 种表现形式，致使其形态变化较为复杂。维吾尔语单词由词干和词缀组成，在同一词干前后添加不同的词缀可以表示不同的词义[14]。由于这些特征，给维吾尔语文本信息处理造成一定的困难，如特征维数大[15]。

所提出的框架由文本处理和分类器两部分组成。图 1 显示了该方法的工作流程。首先，将数字文本保存在 UTF-8 格式的文本文件中。然后进行文本处理，将原始文本转换为特征向量表示。最后，采用提出的改进型 SVM 分类器（CUB-SVM），训练该分类器以建立一个分类模型，用于分类测试样本。

本文方法的其主要创新点为：a)考虑到维吾尔语的特性，本文采用了 N-gram 统计模型进行词干提取，同时采用了一种语义相似性方法来进一步归类词干，减少文本的特征数量和稀疏性;b)为了提高对不平衡数据的分类精度，本文在传统 SVM 基础上开发了一种改进型 SVM。

2　文本预处理与向量表示

通过应用传统技术来分类短文本会产生大量和稀疏的特征向量，从而导致分类器的性能较差。在这项研究中，本文提出一种基于词干的特征约简方法，在文本转换为特征向量之前，应用多个预处理步骤来减少特征向量的稀疏性。这种方法可以分五个阶段来描述，如图 2 所示。

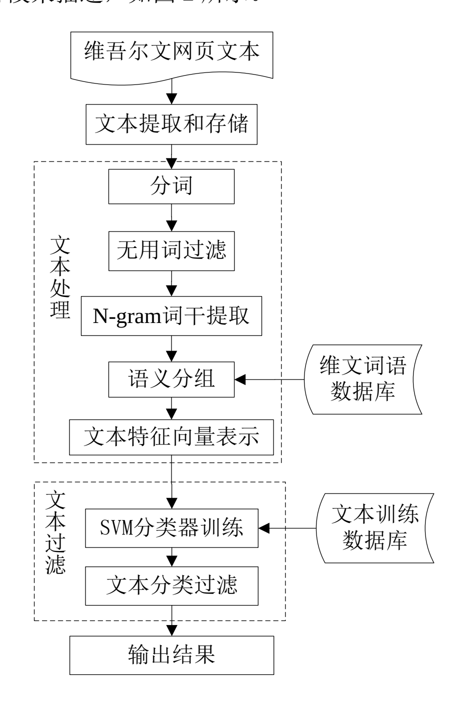


图 1　所提出方法的流程图

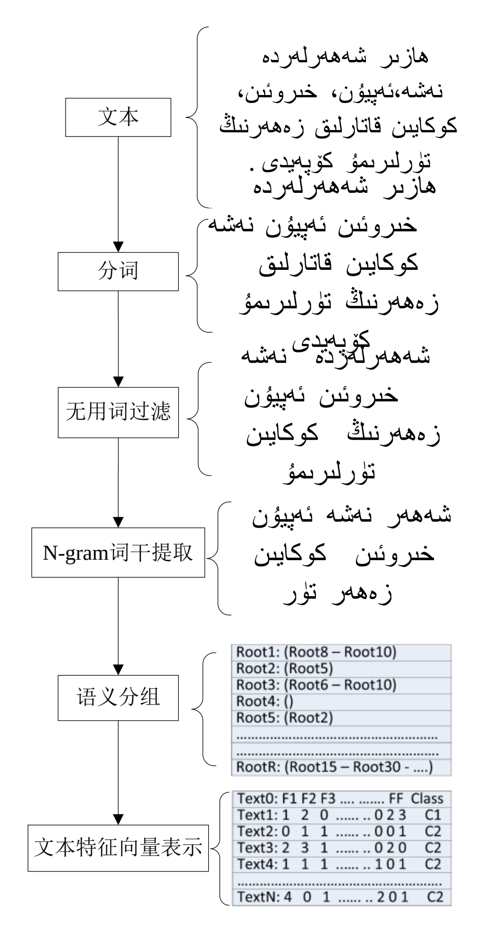


图 2　文本处理与向量表示流程图

1）词语分割

首先，使用斯坦福分词器[16]对原始文本进行分词，将介词

和代词从原词中分离出来，并分隔任何标点符号。

2）无用词过滤

文本无用词过滤过程包括删除停止词、代词、数字、标点和其他非维吾尔语符号。这些部分只会增加特征向量的大小，对区分文本没有帮助。

3）词干提取

维吾尔语是一种以词干为主的语言，这意味着几乎每一个词都是其自身的根源，或者源于三个字母或四个字母的根。从同一根派生的词具有相似的含义，因此可以根据它们的根分组。因此，词干可以被认为是一个特征，从而减少特征向量的长度。

采用更适合维吾尔语环境的统计方法来提取词干。所采用的统计方法为 n-gram 统计模型[17]。在字母层上进行单词切分，即将连续 N 个字母作为一个 gram 单元。n-gram 模型中，对于文本中一个不良字母 l_i，设定其出现的概率与前面 N-1 个字母的出现情况相关。因此，字母序列 $L=l_1l_2l_3,\ldots,l_N$ 出现的概率为：

$$P(L)=P(l_1l_2l_3,\ldots,l_N)=\prod_{i=1}^{N}P(l_i\mid l_{i-N+1},\ldots,l_{i-1}) \tag{1}$$

N-gram 模型中 N 的设定需要结合具体的语言环境，对于维吾尔语，由于其每个单词都由多个字母结合而成，为此较小的 N 不能有效地代表单词属性，而 N 较大如等于 3 或 4 时，则具有较强的代表性。

本文利用 n-gram 统计模型提取词干过程中，为了降低单词维度和冗余度，首先根据维吾尔语词典，删除了单词中最常见的词缀。然后，计算两个词语的相似度[18]，以此来提取词干。

为了展示 n-gram 统计模型提取词语的过程，列举了一个 N=2 时的例子，即计算两个词مائارىپ（教育）和 مائارىپى（教育的）的相似度。

1.مائارىپى ⇒ ما، ئا، رى، پى。（首先将词分解为 N=2 字母组合单元）

2.去除常用词缀的两字母组合 ⇒ ما، ئا، رى، پ。

3.مائارىپ ⇒ ما، ئا، رى،پ。

4.去除常用词缀的两字母组合 ⇒ ما، ئا، رى، پ 。

那么，这两个单词的相似性为：$S=\frac{2C}{A+B}=\frac{2\times 3}{4+3}=0.8571$。

其中，A 表示第一个单词中所包含的且第二个单词中不存在的字母组合的数量；同样，B 第二个单词中所包含的且第一个单词中不存在的字母组合的数量；C 表示两个词中都包含的相同字母组合的数量。若两个单词的相似性大于设定的阈值，则将这两个词合并为一个词干。

4）词义分组

词干有助于将单词与属于同一词干的单词进行分组，但是一些具有相似含义的单词不共享相同的词干。因此本文使用语义方法按以下方式对具有相似含义的词干进行分组。

首先，将来自数据集的每个词干作为查询词，根据同义词词典返回包含该词干的同义词。然后，从文本源中提取同义词并存储在包含词干和其同义词的列表中。最后，将列表中的词干一起比较。如果一个词干与另一个词干共享一个同义词，则这些词干被认为具有相似的含义并被分组在一起。如果一个词干与已在组中的词干共享同义词，则新的词干将添加到现有组中。该过程仅执行一次迭代，这意味着所得到的组不会再聚合在一起。

图 3 展示了语义分组阶段的过程。到这个阶段结束时，具有相似含义的词干被分组在一起，并且可以认为是一个特征。

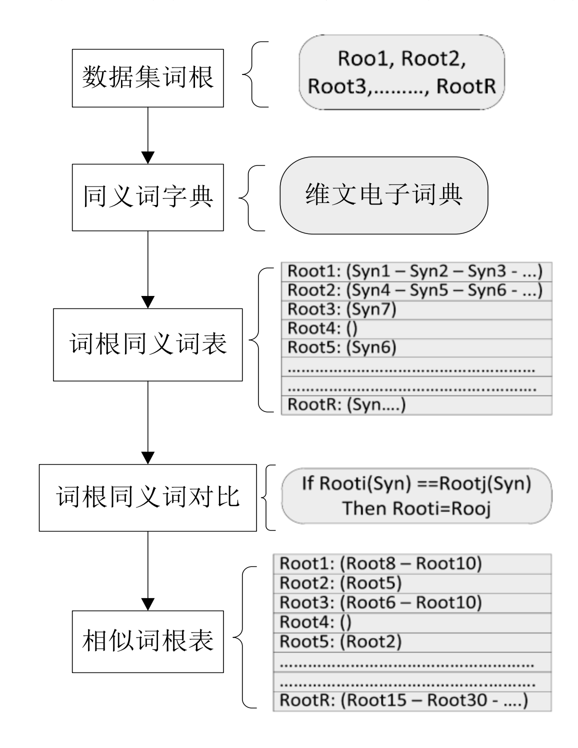


图 3　词义分组示例

5）特征向量构建

在这个阶段，根据获得的词干特征，为数据集中的每个文本构建特征向量，并且整个数据集可以以表格形式呈现。

3　提出的改进型 SVM 文本分类器

3.1　传统 SVM 分类器

SVM 分类器是一种基于统计学的机器学习方法，其基本思想就是通过一个非线性映射，将数据映射到高维特征空间，然后执行线性回归。

设定输入一个数据集 $G=\{(x_i,y_i)\}_{i=1}^{n}$，$x\in R^n$ 为输入特征向量，y 为期望输出值。对于二分类问题，SVM 的映射函数表示为：

$$y=w\phi(x)+b \tag{2}$$

其中：$\phi(x)$ 表示将数据映射到高维空间，参数 w 和 b 的值是通过最小化下式来近似获得：

$$R_{SVM}(c)=(c/n)\sum_{i=1}^{n}L_{\varepsilon}[y_i,w\phi(x_i)+b]+\|w\|^2/2 \tag{3}$$

为了方便计算，引入两个松弛变量 ξ_i^+ 和 ξ_i^-，那么就变成通过最小化下式来估计参数值：

$$R_{SVM}(w,\xi_i^+)=c\sum_{i=1}^{n}(\xi_i^+ +\xi_i^-)+\|w\|^2/2 \tag{4}$$

其中：满足以下条件：

$$w\phi(x_i)+b_i-y_i\le\varepsilon+\xi_i^+ \tag{5}$$

$$y_i-w\phi(x_i)-b_i\le\varepsilon+\xi_i^- \tag{6}$$

3.2 改进的类不平衡 SVM(CUB-SVM)

在这项工作中，为了使 SVM 能够很好地处理具有非线性不可分和不平衡性的维吾尔语文本，本文扩展了传统 SVM 分类器，形成类不平衡 SVM(CUB-SVM)。

除了将内核集成到 SVM 之外，本文引入了一个新的参数 τ_i。它将多数数据样本与分离超平面之间的距离最小化，并将少数数据样本与分离超平面之间的距离最大化，如图 4 所示。CUB-SVM 的目标公式为

$$\min_{w}\left\{\begin{aligned}&\frac{1}{2}\|w\|^2+C^+\sum_{\{i|y_i=+1\}}^{N_1}\xi_i^+ +C^-\sum_{\{i|y_i=-1\}}^{N_2}\xi_i^-\\&+D^+\sum_{\{i|y_i=+1\}}^{N_1}\tau_i^+ -D^-\sum_{\{i|y_i=-1\}}^{N_2}\tau_i^-\end{aligned}\right\} \tag{7}$$

满足：

$$w^T\phi(x_i)+b\ge\tau_i^+-\xi_i^+\ for\ \ x_i{:}y_i=+1 \tag{8}$$

$$w^T\phi(x_i)+b\le-\tau_i^--\xi_i^-\ for\ \ x_i{:}y_i=-1 \tag{9}$$

$$\tau_i^+,\xi_i^+,\tau_i^-,\xi_i^-\ge 0\ for\ \ \forall x_i \tag{10}$$

其中：下标“+”代表多数类别，“-”代表少数类别，并且 C^+,C^-,D^+,D^- 为比例参数。

那么，该问题的拉格朗日方程为

$$\begin{aligned}\mathcal{L}_p=&\frac{1}{2}\|w\|^2+C^+\sum_{\{i|y_i=+1\}}^{N_1}\xi_i^+ +C^-\sum_{\{i|y_i=-1\}}^{N_2}\xi_i^- +D^+\sum_{\{i|y_i=+1\}}^{N_1}\tau_i^+\\&-D^-\sum_{\{i|y_i=-1\}}^{N_2}\tau_i^- -\sum_{\{i|y_i=+1\}}^{N_1}\lambda_i[w^T\phi(x_i)+b-\tau_i^++\xi_i^+]+\\&\sum_{\{i|y_i=-1\}}^{N_2}\mu_i[w^T\phi(x_i)+b+\tau_i^-+\xi_i^-]-\sum_{\{i|y_i=+1\}}^{N_1}\alpha_i\xi_i^+-\\&\sum_{\{i|y_i=-1\}}^{N_2}\beta_i\xi_i^- -\sum_{\{i|y_i=+1\}}^{N_1}\gamma_i\tau_i^+ -\sum_{\{i|y_i=-1\}}^{N_2}\delta_i\tau_i^-\end{aligned} \tag{11}$$

其中：$\lambda_i,\mu_i,\alpha_i,\beta_i,\gamma_i,\delta_i$ 为拉格朗日乘子。通过找到 KKT 条件并代入拉格朗日函数，本文可以得到一个双重问题，表示为。

$$\max_{\lambda,\mu}\left\{\begin{aligned}&\sum_{\{i|y_i=+1\}}^{N_1}\sum_{\{j|y_j=-1\}}^{N_2}\lambda_i\mu_j\phi(x_i)\phi(x_j)\\&-\frac{1}{2}\sum_{\{i|y_i=+1\}}^{N_1}\sum_{\{j|y_j=+1\}}^{N_1}\lambda_i\lambda_j\phi(x_i)\phi(x_j)\\&-\frac{1}{2}\sum_{\{i|y_i=-1\}}^{N_2}\sum_{\{j|y_j=-1\}}^{N_2}\mu_i\mu_j\phi(x_i)\phi(x_j)\end{aligned}\right\} \tag{12}$$

满足：

$$\sum_{\{i|y_i=+1\}}^{N_1}\lambda_i-\sum_{\{i|y_i=-1\}}^{N_2}\mu_i=0 \tag{13}$$

$$0\le\lambda_i\le C^+ \tag{14}$$

$$\mu_i\ge D^- \tag{15}$$

使用 $K(x_i,x_j)=\phi(x_i)\phi(x_j)$，那么 CUB-SVM 可以表示为

$$\max_{\lambda,\mu}\left\{\begin{aligned}&\sum_{\{i|y_i=+1\}}^{N_1}\sum_{\{j|y_j=-1\}}^{N_2}\lambda_i\mu_jK(x_i,x_j)\\&-\frac{1}{2}\sum_{\{i|y_i=+1\}}^{N_1}\sum_{\{j|y_j=+1\}}^{N_1}\lambda_i\lambda_jK(x_i,x_j)\\&-\frac{1}{2}\sum_{\{i|y_i=-1\}}^{N_2}\sum_{\{j|y_j=-1\}}^{N_2}\mu_i\mu_jK(x_i,x_j)\end{aligned}\right\} \tag{16}$$

满足

$$\sum_{\{i|y_i=+1\}}^{N_1}\lambda_i-\sum_{\{i|y_i=-1\}}^{N_2}\mu_i=0 \tag{17}$$

$$0\le\lambda_i\le C^+ \tag{18}$$

$$\mu_i\ge D^- \tag{19}$$

在解决 λ_i,μ_i 的这个问题后，本文可以找到分离超平面，其中：

$$w=\sum_{\{i|y_i=+1\}}^{N_1}\lambda_i\phi(x_i)-\sum_{\{i|y_i=-1\}}^{N_2}\mu_i\phi(x_i) \tag{20}$$

并且，分类器的公式变为

$$f(x)=sign(\sum_{\{i|y_i=+1\}}^{N_1}\lambda_iK(x,x_i)-\sum_{\{i|y_i=-1\}}^{N_2}\mu_iK(x,x_i)+b) \tag{21}$$

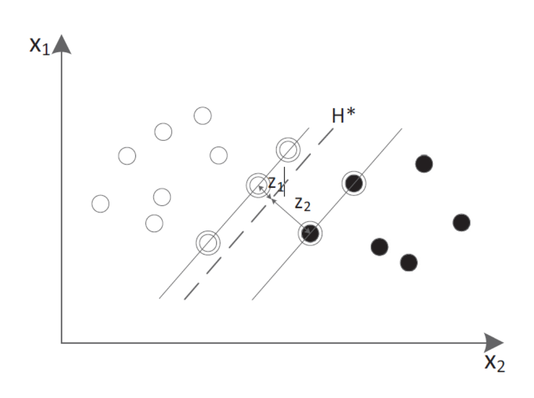


图 4　CUB-SVM 超平面和余量

4　实验及分析

4.1　实验设置

为了测试网页文本中的各种不良主题，从各种维吾尔文网站论坛上收集了 500 篇文本，分为 4 大类：(1)毒品类的文本，数量为 143 篇；(2)色情类的文本，数量为 78 篇；(3)赌博类的文本，数量为 107 篇；(4)正常文本，数量为 172 篇。这些文本及其类别具有不平衡性，文本集的字符长度统计如表 1 所示。

表 1　文本数据集的属性

	最小值	最大值	均值	标准差
字符数	91	2030	834	484
单词数	25	460	195	113

另外，在具有 Intel 酷睿 I5 5250@2.7GHz CPU，24 GB 内存和 Windows7 64 位 PC 机上，通过使用 MATLAB R2013b 和 CVX 2.1 工具箱实现本文方法。

4.2　性能指标

对数据集进行五重交叉验证。为了测量分类器的性能，使用了三个度量标准，其中设定 TP 表示正确分类的阳性样本，

FP 表示错误分类的阳性样本，TN 表示正确分类的阴性样本，FN 表示错误分类的阴性样本。

a)准确性度量(Accuracy)，用于评估分类器正确分类的性能，表示如下。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

b)*F*-measure 度量，用于评估分类器的整体性能。其中，*F*-measure 由精确性(Precision)和召回率(Recall)计算而来，分别表示如下：

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

$$Recall = \frac{TN}{TN + FN} \tag{24}$$

$$F\text{-}Measure = 2 \times (\frac{Precision \times Recall}{Precision + Recall}) \tag{25}$$

c) AUC 面积。感受性曲线(ROC)是以真阳性率和假阳性率为坐标的曲线，曲线与 X 坐标轴之间的面积则为 AUC 面积，取值为 0.5 到 1 之间，用来反映分类器的效果。其值越大说明分类效果越好。

4.3 分类结果分析

首先，本文分析了预处理阶段的效果。在传统文本处理方法中，直接将词语作为特征而不进行过滤和词干化。对于本文采用的维吾尔语文本数据集，传统方法会产生长度为 6204 的特征向量。而本文方法经过词干提取和语义分组后，所获得特征向量长度为 1163，缩小了近 5.3 倍。

为了评估本文 CUB-SVM 的性能，将其与标准 SVM、文献[13]提出的 MINSVM 和文献[12]提出的 SMOTE-SVM 进行比较，以突出不同改进型分类方法之间的区别。表 2 给出了各种方法在数据集上的性能平均值。图 3 给出了各种方法的 ROC 曲线。

表 2　各分类器的平均分类结果

分类器	准确性(%)	精度(%)	召回率(%)	F-度量	AUC 面积
本文 CUB-SVM	0.92	0.87	0.89	0.88	0.94
SVM	0.83	0.73	0.80	0.76	0.83
MINSVM	0.87	0.81	0.85	0.83	0.87
SMOTE-SVM	0.89	0.84	0.86	0.85	0.91

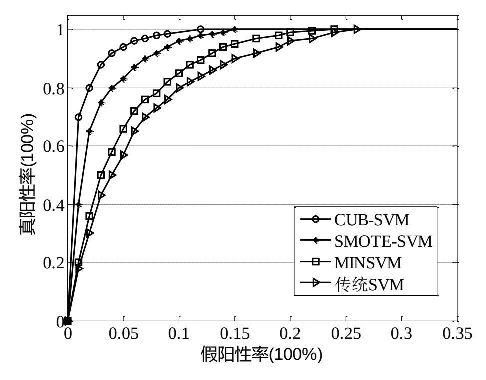


图 3　各种分类方法的 ROC 曲线

可以看出，本文 CUB-SVM 分类器优于传统 SVM 和一些改进 SVM 分类器，其中 *F*-measure 比传统 SVM 提高了 16%。这是因为，本文方法中的预处理过程可以有效降低数据维度，且提出的 CUB-SVM 能够很好地对不平衡短文本进行分类。

对于 SMOTE 和 MINSVM 方法，数据重采样技术并不能提高 SVM 分类器的性能，这是因为其通过改变数据的分布可能导致更多的异常值来降低 SVM 的性能。另外，具有不同成本函数的 SVM 确实提高了分类器的精确性，但是这种改进是以降低召回率为代价的，这导致较低的 F-measure 值和整体准确度。

另外，表 3 给出了各种分类器的分类时间，其中为了公平比较，维吾尔语文本数据的预处理中，都采用了本文提出的过滤和词干提取步骤。可以看出，SMOTE-SVM 具有最高的运行时间，因为它需要对数据执行过采样，这非常耗时且导致产生更多的数据。MINSVM 具有最低的处理时间，因为它随机删除部分数据样本，这导致数据集更小，因此处理时间更短。本文 CUB-SVM 分类器的处理时间比标准 SVM 分类器稍长，但并没有明显增加开销。

表 3　各种分类器的分类时间

分类器	本文 CUB-SVM	SVM	MINSVM	SMOTE-SVM
分类时间(s)	21	16	43	58

4.4 统计分析

统计分析用于测试分类器之间准确度差异的显著性。给定两个分类器，统计测试分类器是否具有相同的期望错误率。为了进行统计分析，本文采用了 K 折交叉验证实验。

K 折交叉验证中，从原始数据集中获得 K 个训练/测试集合对。分类器在训练集 $train_i$ 上训练并在测试集 $test_i$ 上测试。训练和测试集上分类器的误差率分别表示为 p_i^1, p_i^2，$i = 1,2,\ldots,K$。

如果分类器具有相同的错误率，则它们应该具有相同的均值，即它们的平均值的差值应该等于 0。对于 K 交叉验证测试，第 i 个错误率的差异为 $p_i = p_i^1 - p_i^2$，为此可得到一个包含 K 点的 p_i 分布。假设 p_i^1 和 p_i^2 都是正态分布的，那么它们的差异 p_i 也是正态分布的。

假设 H_0 为这种分布具有正常的零均值。

$$H_0 : \mu = 0, \quad H_1 : \mu \neq 0$$

$$Let : S = \frac{\sum_{i=1}^{K} p_i}{K}, S^2 = \frac{\sum_{i=1}^{K} (p_i - m)^2}{K - 1}$$

在 $\mu = 0$ 的零假设下，有一个符合 t-分布的自由度为 $K-1$ 的统计量：$\frac{\sqrt{K}.m}{S} \sim t_{k-1}$ 。

如果该值在范围 $(-t_{\alpha/2,K-1}, -t_{\alpha/2,K-1})$ 之外，则测试拒绝在显著性水平 α 上的假设。当 $\alpha = 0.1$ 时，置信水平为 90% ，范围为（-2.132，2.132）。

测试的错误率以及整体错误率的统计分析结果如表 4 所示。由于数据集具有不平衡性，有些类别的样本数量较少。统计结果可以看出，所有测试均拒绝少数类别的假设。对于多数

类别来说，假设在两个数据集上被拒绝并在其他三个数据集上被接受，这意味着两个数据集中的错误率在其他分类器上存在差异，但在其他三个数据集上没有差异。至于整体错误率，所有数据都接受假设，这意味着整体错误率没有差异。总之，本文 CUB-SVM 分类器在少数类别上具有较好的准确性，同时不会牺牲整体的准确性，这说明了本文分类器能够很好地处理不平衡数据。

表 4　分类器成对统计分析的分数

分类器对	多数类分数	少数类分数	总体分数
CUB-SVM vs. SVM	0.561	−2.953	−1.452
	accepted	accepted	accepted
CUB-SVM vs. SMOTE-SVM	1.742	−4.765	−0.503
	rejected	accepted	accepted
CUB-SVM vs. MINSVM	0.405	−2.526	−1.416
	rejected	accepted	accepted

5　结束语

本文提出了一种用于维吾尔语网站文本过滤的方法，采用了不良的文本预处理和词干提取步骤来降低文本特征维度，并将本文向量化表示。为了更好的对类不平衡文本进行分类，提出了一种改进型的 SVM 分类器(CUB-SVM)，以此实现维吾尔语文本的高精度分类，从而对不良文本进行过滤。实验结果表明，提出的方法能够准确分类出不良类的网页文本，能够应用于维吾尔语网页的管理和净化。

参考文献：

[1] 余本功，张连彬. 基于 CP-CNN 的中文短文本分类研究 [J]. 计算机应用研究, 2018, 35 (4): 1001-1004. (Yu Bengong, Zhang Lianbin. Chinese short text classification based on CP-CNN [J]. Application Research of Computers, 2018, 35 (4): 1001-1004.)

[2] 黄发良，冯时，王大玲，等. 基于多特征融合的微博主题情感挖掘 [J]. 计算机学报，2017，40 (4): 872-888. (Huang Faliang, Fneg Shi, Wang Daling, *et al.* Mining topic sentiment in microblogging based on multi-feature fusion [J]. Chinese Journal of Computers, 2017, 40 (4): 872-888.)

[3] 顾晓清，蒋亦樟，王士同. 用于不平衡数据分类的 0 阶 TSK 型模糊系统 [J]. 自动化学报, 2017, 43 (10): 1773-1787. (Gu Xiaoqing, Jiang Yizhang, Wang Shitong. Zero-order TSK-type fuzzy system for imbalanced data classification [J]. Acta Automatica Sinica, 2017, 43 (10): 1773-1787.)

[4] Bhowan U, Mark Johnston, Zhang Mengjie, *et al.* Evolving diverse ensembles using genetic programming for classification with unbalanced data [J]. IEEE Trans on Evolutionary Computation, 2013, 17 (3): 368-386.

[5] 买买提依明•哈斯木，吾守尔•斯拉木，维尼拉•木沙江，等. 基于 N 元模型的维吾尔语文本分类技术研究 [J]. 计算机应用研究，2015, 32 (7): 1986-1988. (Maimaitiyiming Hasimu, Wushouer Silamu, Weinila Mushajiang Nuermaimaiti Youluwasi, *et al.* Research N-gram based Uyghur text classification technique [J]. Application Research of Computers, 2015, 32 (7): 1986-1988.)

[6] 吐尔地•托合提，艾克白尔•帕塔尔，艾斯卡尔•艾木都拉. 语义词特征提取及其在维吾尔文文本分类中的应用 [J]. 中文信息学报，2014, 28 (6): 140-144. (Turdi Tohti, Akbar Pattar, Askar Hamdulla. Semantics-based feature extraction and its application in uyghur text classification [J]. Journal of Chinese Information Processing, 2014, 28 (6): 140-144.)

[7] 吐尔地•托合提，艾海麦提江•阿布来提，米也塞•艾尼玩，等. 一种结合 GAAC 和 K-means 的维吾尔文文本聚类算法 [J]. 计算机工程与科学, 2013, 35 (7): 149-155. (Turdi Tohti, Ahmatjan Ablat, Muyassar Aniwar, *et al.* Combined algorithm of GAAC and K-means for Uyghur text clustering [J]. Computer Engineering and Science, 2013, 35 (7): 149-155.)

[8] 李响，吐尔根•依布拉音，卡哈尔江•阿比的热西提，等. 基于主动学习的 SVM 维吾尔语情感分析研究 [J]. 新疆大学学报：自然科学版, 2015, 32 (4): 447-452. (LI Xiang, Tuergen Yibulayin, Kahaerjiang Abiderexiti, *et al.* Emotion analysis of active learning based on SVM in Uyghur language [J]. Journal of Xinjiang University (Natural Science Edition) , 2015, 32 (4): 447-452.)

[9] Pal M, Mather P M. Support vector machines for classification in remote sensing [J]. International Journal of Remote Sensing, 2015, 26 (5): 1007-1011.

[10] Yuan Yubo, Fan Weiguo, Pu Dongmei. Spline function smooth support vector machine for classification [J]. Journal of Industrial & Management Optimization, 2017, 3 (3): 529-542.

[11] Ma Hongyan, Wang Liling, Shen Bo. A new fuzzy support vector machines for class imbalance learning [C]// Proc of International Conference on Electrical and Control Engineering. Piscataway, NJ: IEEE, 2011: 3781-3784.

[12] Chawla N V, Bowyer K W, Hall L O, *et al.* SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16 (1): 321-357.

[13] Ajeeb N, Nayal A, Awad M. Minority SVM for linearly separable imbalanced datasets [C]// Proc of International Joint Conference on Neural Networks. : Piscataway, NJ: IEEE, 2014: 1-5.

[14] 阿不都萨拉木•达吾提，于斯音•于苏普，艾斯卡尔•艾木都拉. 类别区分词与情感词典相结合的维吾尔文句子情感分类 [J]. 清华大学学报：自然科学版，2017, 57 (2): 197-201. (Abdusalam Dawut, Hussein Yusuf, Askar Handulla. Emotion recognition from Uyghur sentences based on combinations of class discrimination words and a sentiment dictionary [J]. Journal of Tsinghua University: Science and Technology, 2017, 57 (2): 197-201.)

[15] 韩军兵，哈力旦・阿布都热依木，古力努尔・艾尔肯，等. 改进信息增益的维吾尔文特征选择方法 [J]. 计算机工程与应用，2017, 53 (23): 34-38. (Han Junbing, Halidan ・ Abudureyimu, Gulnur Arken, *et al.* Improved information gain algorithm based on Uyghur feature selection [J]. Computer Engineering and Applications, 2017, 53 (23): 34-38.)

[16] Sun Rong, Zhou Wen, Liu Zongtian. Using language rules to improve the performance of word segmentation [C]// Proc of International Congress on Image and Signal Processing. Piscataway, NJ: IEEE, 2014: 1665-1669.

[17] 于洁. 基于 Spark 和 DN-gram 模型的定义抽取研究 [J]. 北京信息科技大学学报：自然科学版，2017, 32 (4): 64-68. (Yu Jie. Research on definition extraction based on Spark and DN-gram model [J]. Journal of Beijing Information Science & Technology University, 2017, 32 (4): 64-68.)

[18] 董洋溢，李伟华，于会. 基于混合余弦相似度的中文文本层次关系挖掘 [J]. 计算机应用研究, 2017, 34 (5): 1406-1409. (Dong Yangyi, Li Weihua, Yu Hui. Hierarchical relation mining of Chinese text based on mixed cosine similarity [J]. Application Research of Computers, 2017, 34 (5): 1406-1409.)